# Open Access aspects at the CERN Document Server

Alberto Pepe and Tibor Simko

1 June 2005

# Welcome to CERN!

**In its 50 years of activity, CERN has been in the frontline in the field of particle physics...**

- world’s largest particle physics centre (20 member states, 3000 employees, 6500 visiting scientists)
- world’s largest particle accelerator
- Nobel prize winners and laureates

**but also...**

- invention of the World Wide Web in 1989
- inspiration for best-selling fiction books!

# CERN's scholarly output

Research in particle physics and related areas results in:

- 2000 scientific publications per year (OA1)
- 10000 conference talks and contributions per year (OA2)
- TONS of experimental data (OA3)

The fundamental conCERN at CERN is:

- **long-term preservation**
- **dissemination**

Large and increasing amount of produced scholarly material
⟶ need for interoperable institutional repositories
⟶ *the arsenals of knowledge!*

# CERN Document Server: A bit of history

- **pre-1993** - paper dissemination of preprints by the CERN Library
- **1993** - CERN Preprint Server on the web
  - institutional repository
  - two collections: CERN preprints, SCAN series
- **1996** - CERN Web Library
  - adding books, periodicals, and other library objects
- **1999** - CERN Agenda
  - sister application for conferences, meetings and workshops
- **2000** - CERN Document Server
  - adding multimedia material (photos, posters, brochures, videos)

# CERN Document Server in 2005

**Integrated Digital Library:** (OA1)

- articles, preprints, books, photos, videos, exhibition objects...
- 800,000 documents
- 60,000 new acquisitions per year (about 1500 direct author submissions)
- 500 collections

**Integrated Digital Conference:** (OA2)

- conference talks, workshop presentations, meeting minutes...
- 90,000 documents
- 10,000 new acquisitions per year
- 15,000 events

# Integrated Digital Library Software

CDSware

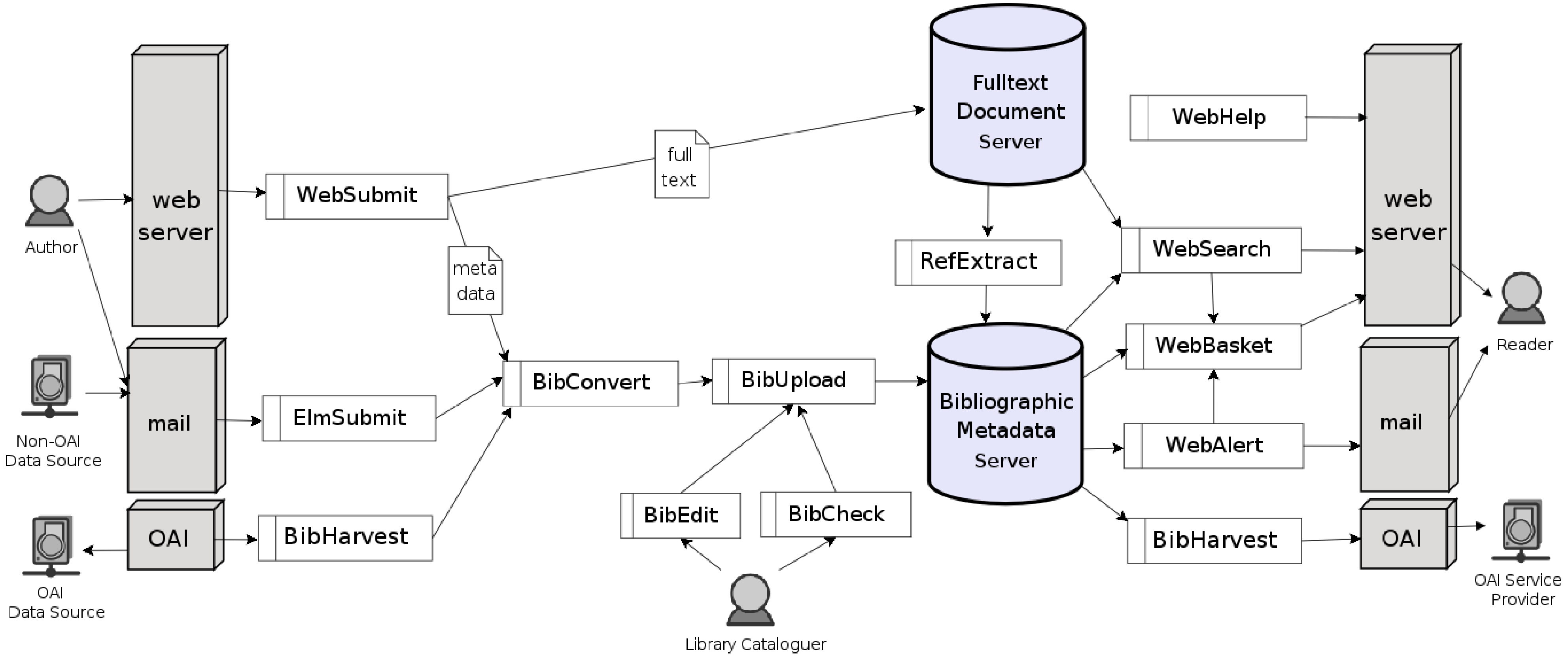
- configurable portal-like interfaces for hosting various kind of collections
- powerful search engine with Google-like syntax
- extensible metadata representation (MARC XML) to handle virtually any kind of document
- flexible document type submission and approbation workflow
- user personalization, including document baskets and email alerts
- I18N, search interface available in 14 languages
- compliant to Open Archive Initiative protocol for metadata harvesting
- co-developed with EPFL, Lausanne
- free software (GNU GPL)

# Integrated Digital Conference Software

INDICO

- handles various kinds of events (lectures, meetings, conferences)
- covers full organizational cycle from registration to creation of proceedings
- user-customized views
- fine access control mechanism
- multiple outputs (HTML, XML, PDF, iCAL, OAI)
- EU project InDiCo (2001-2004)
- free software (GNU GPL)

# CDSware: from data acquisition to delivery (OA1)



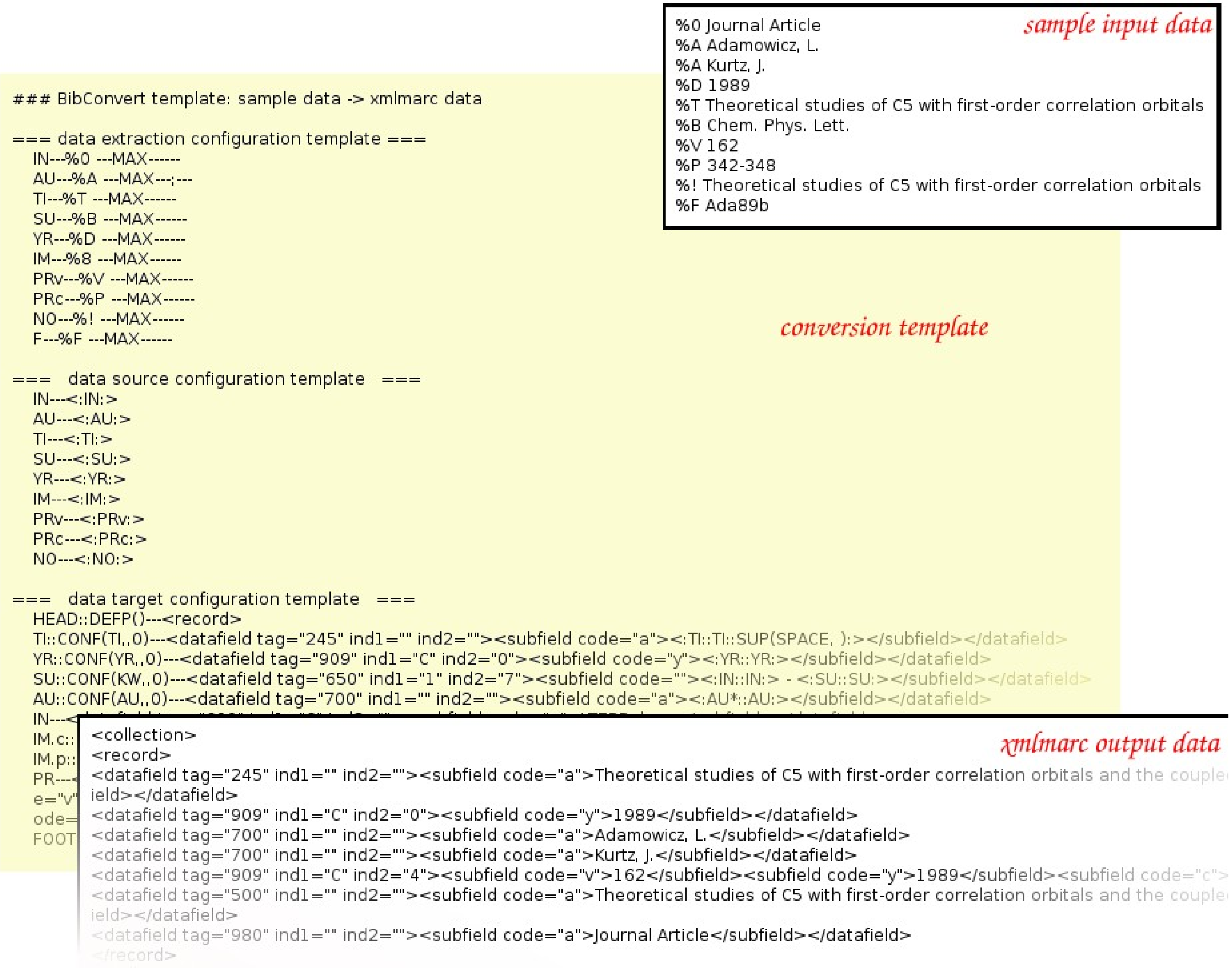
# CDS Metadata Provider: OAI gateway (OA1)

- all CERN-produced documents exposed through the OAI-PMH
- about 40,000 records available
- about 30,000 harvesting requests per month in 2004
  (only 5,000 in 2003!)
- metadata formats `oai_dc`, `marcxml`
- need to improve recognition of papers by locally available ranking methods:
  - ranking by number of citations
  - ranking by downloads and by downloaders
  - ... and even ranking by journal impact factors
- exporting of enriched external records.. added-value by:
  (i) CERN Library (e.g. conference information)
  (ii) automatic processing (e.g. keywords and citations)

# CDS Service Provider: Automated Metadata Harvesting (OA1)

- about 60,000 new acquisitions harvested per year
- at present, CDS harvests metadata from more than 100 sources:
  - only about 2-3 sources are OAI-compliant(!)
  - ... but the most important source is (arXiv.org, about 70% of import volume traffic)
- current harvesting mechanism relies on arXiv.org email submission system
  - richer metadata content
  - ... but non-OAI
- need richer metadata format, current `oai_dc` is very spartan
- need more OAI-compliant data sources

# Example of Automated Metadata Harvesting (OA1)



sample input data

```
%0 Journal Article
%A Adamowicz, L.
%A Kurtz, J.
%D 1989
%T Theoretical studies of C5 with first-order correlation orbitals
%B Chem. Phys. Lett.
%V 162
%P 342-348
%! Theoretical studies of C5 with first-order correlation orbitals
%F Ada89b
```

conversion template

```
### BibConvert template: sample data -> xmlmarc data

=== data extraction configuration template ===
   IN---%0 ---MAX------
   AU---%A ---MAX---;---
   TI---%T ---MAX------
   SU---%B ---MAX------
   YR---%D ---MAX------
   IM---%8 ---MAX------
   PRv---%V ---MAX------
   PRc---%P ---MAX------
   NO---%! ---MAX------
   F---%F ---MAX------

===   data source configuration template   ===
   IN---<:IN:>
   AU---<:AU:>
   TI---<:TI:>
   SU---<:SU:>
   YR---<:YR:>
   IM---<:IM:>
   PRv---<:PRv:>
   PRc---<:PRc:>
   NO---<:NO:>

===   data target configuration template   ===
   HEAD::DEFP()---<record>
   TI::CONF(TI,,0)---<datafield tag="245" ind1="" ind2=""><subfield code="a"><:TI::TI::SUP(SPACE, ):></subfield></datafield>
   YR::CONF(YR,,0)---<datafield tag="909" ind1="C" ind2="0"><subfield code="y"><:YR::YR:></subfield></datafield>
   SU::CONF(KW,,0)---<datafield tag="650" ind1="1" ind2="7"><subfield code=""><:IN::IN:> - <:SU::SU:></subfield></datafield>
   AU::CONF(AU,,0)---<datafield tag="700" ind1="" ind2=""><subfield code="a"><:AU*::AU:></subfield></datafield>
   IN---<
   IM.c::
   IM.p::
   PR---<
   e="v"
   ode=
   FOOT
```

xmlmarc output data

```
<collection>
<record>
<datafield tag="245" ind1="" ind2=""><subfield code="a">Theoretical studies of C5 with first-order correlation orbitals and the couple
ield></datafield>
<datafield tag="909" ind1="C" ind2="0"><subfield code="y">1989</subfield></datafield>
<datafield tag="700" ind1="" ind2=""><subfield code="a">Adamowicz, L.</subfield></datafield>
<datafield tag="700" ind1="" ind2=""><subfield code="a">Kurtz, J.</subfield></datafield>
<datafield tag="909" ind1="C" ind2="4"><subfield code="v">162</subfield><subfield code="y">1989</subfield><subfield code="c">
<datafield tag="500" ind1="" ind2=""><subfield code="a">Theoretical studies of C5 with first-order correlation orbitals and the couple
ield></datafield>
<datafield tag="980" ind1="" ind2=""><subfield code="a">Journal Article</subfield></datafield>
</record>
```

# Conferencing: Enforce institutional self-archiving (OA2)

**Ongoing goal: ensuring Open Access to conference material presented by CERN authors:**

- ensure fast dissemination of conference contributions through OAI
- encourage paper submission within the CERN administrative procedures for travel request
- promote the use of OAI-compliant conference management software

**JISC 2004 and 2005 self-archiving survey:**

- most researches don’t self-archive and won’t, unless required by employer
- when required, 81% will comply willingly, 14% reluctantly, 5% not at all
- good, but...

# Sharing Raw Research Data (OA3)

- *“archives containing hundreds or thousands of terabytes of data will be affordable and necessary for archiving scientific and engineering information”*
- LHC will produce huge amounts of raw experimental data
- current technology is fine to provide OA to relatively large collections of documents, but..
- need a widely accepted, solid, reliable infrastructure to allow global collaboration
- active successful projects in Astrophysics (Virtual Observatory) and Chemistry (Comb-*e*-Chem)
- in particle physics? CERN is paving the road for a common infrastructure to allow data and resource sharing on a global scale... the Grid!

# Conclusions

- CDS: more than 10 years of experience in handling digital documents
- CDS Software for Open Access
  - CDSware: Integrated Digital Library
  - InDiCo: Integrated Digital Conference
- need for detailed metadata description (MARC)
- need for interoperability “beyond” OAI
- need for fast dissemination of conference contribution
- actively promote institutional self-archiving
- Open Access to data?